ARL-TR-8217 ● Nov 2017

US Army Research Laboratory

Application of the Fractions Skill Score for Tracking the Effectiveness of Improvements Made to Weather Research and Forecasting Model Simulations

by John W Raby and Huaqing Cai

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8217 ● Nov 2017

US Army Research Laboratory

Application of the Fractions Skill Score for Tracking the Effectiveness of Improvements Made to Weather Research and Forecasting Model Simulations

by John W Raby and Huaqing Cai
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
November 2017	Technical Report	October 2016–August 2017

4. TITLE AND SUBTITLE
Application of the Fractions Skill Score for Tracking the Effectiveness of Improvements Made to Weather Research and Forecasting Model Simulations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
John W Raby and Huaqing Cai

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
Computational and Information Sciences Directorate
ATTN: RDRL-CIE-M
White Sands Missile Range, NM 88002

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8217

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Spatial forecasts from Numerical Weather Prediction (NWP) models of meteorological variables supporting US Army battlefield operations are an integral part of the products available for the Staff Weather Officer's use in providing mission forecasts. This report presents some preliminary results obtained from the application of a nontraditional fuzzy verification method to evaluate the ability of NWP to simulate spatial variable fields filtered using thresholds. Fuzzy methods have been developed in recent years to overcome limitations encountered when applying traditional verification techniques to high-resolution NWP forecasts, which often result in misleading assessments of forecast accuracy. This study illustrates how the Fractions Skill Score (FSS) generated by the Model Evaluation Tools can be applied to assess the US Army Research Laboratory's Weather Running Estimate–Nowcast (WRE–N) model forecasts. The FSS is widely recognized as an important metric for verifying model performance as a function of threshold value and spatial scale and, when used to characterize the baseline performance, provides the basis for comparison as model improvements are implemented. Preliminary results suggest that the FSS applied to assess the WRE–N provides a robust metric to track changes in model performance and a better metric of the skill in predicting objects that affect input to My Weather Impact Decision Aid.

15. SUBJECT TERMS
fuzzy, neighborhood, spatial scale, thresholds, numerical weather prediction, observations, model performance, model evaluation tools, fractions skill score, FSS, assessment, My Weather Impact Decision Aid, MyWIDA, weather impacts

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John W Raby
Unclassified	Unclassified	Unclassified	UU	40	19b. TELEPHONE NUMBER (Include area code) 575-678-2004

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Preface

This technical report relates to a previous work that explored the application of fuzzy verification methods to verify spatial forecasts produced by the Weather Running Estimate–Nowcast (WRE–N) of continuous meteorological variables that have been filtered by a threshold. These methods use gridded forecasts and observations on a common grid, which enabled the application of a number of different spatial verification methods that reveal various aspects of model performance. This report describes how the Fractions Skill Score can be used to assess the performance of the WRE–N in predicting objects and to track the changes in performance attributable to model upgrades. Portions of this report's content originated in ARL-TR-7849.[1]

[1]Raby JW. Application of a fuzzy verification technique for assessment of the weather running estimate−nowcast (WRE−N). White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Aug. Report No.: ARL-TR-7849.

Acknowledgments

We offer our thanks to Mr Robert Dumais of the US Army Research Laboratory (ARL) who contributed guidance, data, suggestions, and information without which the study could not have been completed. We also thank Dr Brian Reen and Dr Jeffrey Smith for their many suggestions for improvements to the application of the fuzzy verification techniques. Many thanks to Mr Mark Gatlin of the ARL Technical Publishing office for his many helpful suggestions that significantly improved the report.

Summary

Spatial forecasts from Numerical Weather Prediction (NWP) models to support the US Army's battlefield operations have become an integral part of the products available for the Air Force Staff Weather Officer to use in providing mission planning and execution forecasts. Tactical decision aids (TDAs) ingest these forecasts. The TDAs fuse information on the characteristic operational weather thresholds that affect the performance of Army systems and missions with the NWP's spatial forecast information to generate spatial forecasts of these impacts for user-specified systems and/or missions for the time period and location of interest. This report presents methods that can be used to verify spatial-forecast fields of meteorological variables that have been filtered by the application of a threshold the same way as that used by the TDA. In effect, thresholds applied to a continuous variable field become categorical forecasts for which there are traditional and nontraditional methods for verification. This study evaluates the applicability of the Fractions Skill Score (FSS), which is computed using a nontraditional, fuzzy verification technique to assess unique aspects of model performance at domain-level in a way traditional techniques alone cannot. The FSS can be used to establish a baseline performance for a range of threshold values and spatial scales. As model upgrades are implemented, the FSS can be used as a metric to track changes in model performance to determine the effectiveness of the changes.

Traditional grid-to-point methods can verify the skill of NWP in predicting continuous meteorological variables by computing such statistics as mean error and root-mean-square error, which characterize model accuracy over the entire domain. When these techniques are applied to high-resolution models such as the Army Weather Running Estimate–Nowcast (WRE–N), the results can give misleading error estimates when compared with lower-resolution models, which often score better when using these techniques. The issue is the inability of the verification technique to evaluate the true skill of higher-resolution forecasts, which replicate mesoscale atmospheric features in a way that is more representative of the actual phenomenon, owing to their use of a reduced grid spacing over smaller domains, higher-resolution land-surface models, and better parameterization of subgrid physical processes.

In recent years, various nontraditional verification techniques have attempted to use different approaches to show the value of higher-resolution forecasts. In particular, spatial verification techniques have been developed that overcome the limitations of grid-to-point techniques, which score on the basis of the exact matching between

point observations and the forecasts at those points. Fuzzy verification, also known as neighborhood verification, uses an approach that does not require exact matching and instead focuses on how well the atmospheric feature or object is replicated by the model, even if there is a spatial displacement of the feature. The goal is to determine the amount of displacement by using a range of sizes of neighborhoods of surrounding forecast and observed grid points in the verification process. In this way, model performance as a function of spatial scale can be determined to allow selection of the scale required to have the desired accuracy. Many methods for fuzzy verification have been developed, mostly for evaluating model precipitation forecasts. Ebert reviews a number of such methods.[1] For this study, the FSS method was applied to continuous meteorological variables.

The fuzzy verification framework used requires the use of observations on a grid matching that of the WRE–N. Neighborhoods are defined in terms of the grid boxes within both grids. The number of grid boxes in one direction determines the size of the 2-D neighborhood or spatial scale when expressed in terms of the grid spacing for a particular model grid. For this study, various spatial scales were chosen. The metrics and diagnostics used for scoring arise by defining an event from both the forecast and the observation grids. The event is defined by the use of a category or threshold as the basis for determining "hits" or "misses", which follows the established theoretical framework for evaluating deterministic binary forecasts. This framework evaluates the forecast skill by counting the numbers of times the event was forecasted—or not—and observed—or not—in a contingency table. There are many statistics and skill scores that can be computed from the data collected by this method. The FSS, developed by Roberts and Lean, uses the fuzzy verification framework and compares the fractional coverage of events in the neighborhoods in the observation grid and the forecast grid to evaluate the closeness of the forecast to the observation.[2]

The authors obtained model output from the Army WRE–N, which is an advanced research version of the Weather Research and Forecasting model adapted for generating short-range nowcasts and gridded observations produced by the National Oceanographic and Atmospheric Administration's Global Systems Division using the Local Analysis and Prediction System. A tool developed by the National Center for Atmospheric Research called Model Evaluation Tools Grid-Stat was used to apply the neighborhoods and thresholds to the grids and

[1]Ebert E. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. Meteo App. 2008;15:51–64.

[2]Roberts NM, Lean HW. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review. 2008;136:78–97.

calculate the aggregate fuzzy-verification skill scores and contingency-table statistics for the entire domain.

Preliminary results suggest that the FSS offers an assessment of model performance in terms of spatial scale and threshold value that promises to provide a metric that can be used to track the effectiveness of model upgrades and to gain insight into the scales and threshold values for which the model shows skill. In addition, it provides a useful metric to assess the ability of the model for prediction of objects defined by thresholds that impacts the input data used by the My Weather Impact Decision Aid.

INTENTIONALLY LEFT BLANK.

1. Introduction

As computing technology has advanced, the weather-forecasting task—once the job of a human forecaster in theater—has shifted to computerized Numerical Weather Prediction (NWP) models. Scientists around the world have used the Weather Research and Forecasting (WRF) model extensively for many applications. In this study, we have used the Advanced Research version of WRF (Skamarock et al. 2008), which we abbreviate as WRF–ARW. WRF–ARW includes Four-Dimensional Data Assimilation (FDDA) techniques that can incorporate observations into the model so that forecast quality is improved (Stauffer and Seaman 1994; Deng et al. 2009). The US Army Research Laboratory (ARL) uses WRF–ARW as the core of its Weather Running Estimate–Nowcast (WRE–N) weather forecasting model.

The Army requires high-resolution weather forecasting to model atmospheric features with wavelengths on the order of 5 km or less, which imposes a requirement for NWP to operate on a model grid spacing on the order of 1 km or less in the finest, or most resolved, domain to resolve weather phenomena of interest to the Soldier in theater. The atmospheric flows of interest to the Army include mountain/valley breezes, sea breezes, and other flows induced by differences in land-surface characteristics. High-resolution NWP forecasts need to be validated against observations before their outputs can be used by applications such as My Weather Impacts Decision Aid (MyWIDA), developed by Brandt et al. (2013). Weather forecast validation has always been of interest to the civilian and military weather forecasting community; see, for example, the reviews by Casati et al. (2008), Ebert et al. (2013), and/or the books by Wilks (2011) or Jolliffe and Stephenson (2012). The validation of the models, especially high-resolution NWP, has proven to be especially difficult when addressing small temporal and spatial scales (NRC 2010) that characterize NWP for use in Army applications. Furthermore, verification has not been accomplished for WRE–N spatial fields of continuous meteorological variables that have been filtered by the application of a threshold to evaluate the applicability of such output for use in MyWIDA.

The WRF model is maintained by the National Center for Atmospheric Research (NCAR), which has also developed a suite of Model Evaluation Tools (MET) (NCAR 2013) to evaluate WRF–ARW performance. MET was developed at NCAR with a grant from the US Air Force 557th Weather Wing (formerly the Air Force Weather Agency) and the National Oceanic and Atmospheric Administration (NOAA). NCAR is sponsored by the United States National Science Foundation. Grid-Stat, a tool in MET, provides verification statistics for matched forecast and

observation grids. For fuzzy or neighborhood verification, Grid-Stat compares the forecasts and observations at grid points in a neighborhood surrounding the point of interest rather than comparing a single point from both fields. For scoring, thresholds are set that establish the basis for defining forecast and observed events. Within a neighborhood, the fractional coverage of observed events and forecast events are determined. By varying the neighborhood size, and rescoring, the relationship between neighborhood size and forecast skill can be determined. Grid-Stat performs this scoring at every grid point to calculate aggregate fuzzy-verification skill scores and contingency-table statistics for the entire domain. This study uses the Fractions Skill Score (FSS), which compares the proportion of grid boxes within a forecast neighborhood that have events with the proportion of grid boxes within the observed neighborhood that have events and results in a score that expresses the skill of the forecast by application of the assumption that useful forecasts are those whose frequency of forecast events is close to that of the observed events (Ebert 2008).

ARL has employed the spatial verification tool in MET called Method for Object-based Diagnostic Evaluation (MODE) in prior assessments such as that of Cai and Dumais (2015). They evaluated the 3-km grid spacing High Resolution Rapid Refresh model to demonstrate the utility of a nontraditional object-based technique in providing additional information to improve model precipitation forecasts to complement the information provided by the use of traditional verification techniques. In a separate study, Vaucher and Raby (2014) developed the capability to use MODE for object-based assessment of 1-km grid spacing WRE–N output of continuous meteorological variables. For this study, the only source of gridded observations available was from the NOAA–National Centers for Environmental Prediction (NCEP) Real-Time Mesoscale Analysis (RTMA) product (De Pondeca et al. 2011). In Vaucher and Raby (2014), the RTMA product, generated at a horizontal grid spacing of 2.5 km, was used with the WRE–N output that was remapped from a 1-km grid to a 2.5-km grid to produce the required matching grid.

MODE proved to be useful as an assessment tool for the WRE–N over an Army-scale domain, and plans were made to expand its use to perform evaluations of continuous meteorological variables generated by the WRE–N at 1.75-km grid spacing. Collaborations with NOAA's Global Systems Division (GSD) resulted in the generation of 1.75-km grids of observations of surface meteorological variables for the same domain as the WRE–N using the NOAA–GSD Local Analysis and Prediction System (LAPS).

MET Series-Analysis was used in combination with MODE to perform spatial verification of the 1.75-km WRE–N by Raby and Cai (2016). This assessment demonstrated the value of combining the results from traditional categorical and nontraditional object-based verification methods for verification of the WRE–N. It also demonstrated how these methods verify spatial forecasts of continuous meteorological variables, which have been filtered by a single-threshold to quantify the degree of this particular type of accuracy, applicable to forecasts being used by the MyWIDA Tactical Decision Aid (TDA).

For this study, the WRE–N was run with FDDA for 5 case study days over a 1.75-km grid-spacing domain in Southern California over highly varied terrain and with a dense observational network that provided a robust data set of model output for analysis. The case study days from February through March 2012 were picked to vary weather conditions from a strong synoptic-forcing situation to a quiescent situation. (The weather conditions for each study day are described in Section 2.3).

This study illustrates how fuzzy verification techniques available from the MET Grid-Stat tool can be applied to the assessment of high-resolution WRE–N model forecasts. Fuzzy verification is a type of spatial verification developed in recent years to address the inability of traditional verification techniques to adequately verify model forecasts, which are generated on increasingly smaller grids to resolve smaller-scale atmospheric features that are of interest to the Army. Traditional grid-to-point techniques score on the basis of the exact match between point observations and the forecasts at those points. When these techniques are applied to forecasts on grids with smaller spacing between grid points, the results are often misleading due to the error statistics being higher than those generated when the same technique is used for forecasts on grids whose points are spaced wider apart. In fact, the atmospheric features replicated by models using small-grid spacing bear more resemblance to the actual phenomena than those simulated at larger-grid spacings.

The problem lies in requiring the exact match between the point observations and the forecast grid values. This leads to the so-called “double penalty” where the feature in the forecast being spatially displaced creates an offset in position that produces 2 types of errors. The first type results from the forecast placing the feature where it was not observed; the second type results from the forecast not placing the feature where it was observed (Mittermaier et al. 2013), Furthermore, the error statistics provide no information about occurrences of “near-misses” that suggest a forecast of some quality or occurrences of more complete misses owing to a poor forecast. The challenge is to employ techniques that evaluate the ability of the model to replicate the features themselves, albeit with displacement, in

addition to the more traditional objective approaches. To this end, researchers have developed spatial-verification techniques that reveal more about the ability of the model to predict spatial features (Jolliffe and Stephenson 2012).

Fuzzy verification uses an approach that does not require exact matching but instead focuses on how well the atmospheric feature or object is replicated by the model, even if there is a spatial displacement of the feature. The goal is to determine the amount of displacement by using a range of sizes of neighborhoods of surrounding forecast and observed grid points in the verification process. In this way, model performance as a function of spatial scale can be determined to allow selection of the scale required to have the desired accuracy. Many methods for fuzzy verification have been developed, mostly for evaluating model-precipitation forecasts. Ebert (2008) reviews a number of such methods. For this study, the FSS method was applied to continuous meteorological variables. Furthermore, Mittermaier et al. (2013) pointed out the difficulties of assessing the true skill of the forecast using traditional categorical verification because of the sensitivity of those statistics to the base rate or observed event frequency that often results in the perception of decreased skill with increasing threshold value.

The FSS method of fuzzy verification used for this study requires the use of observations on a grid matching that of the WRE–N. Neighborhoods are defined in terms of the grid boxes within both grids. The number of grid boxes in one direction determines the size of the 2-D neighborhood or spatial scale when expressed in terms of the grid spacing for a particular model grid. For this study, various spatial scales were chosen. The metrics and diagnostics used for scoring arise by defining an event from both the forecast and the observation grids. The event is defined by the use of a category or threshold that serves as the basis for determining “hits” or “misses”, which follows the established theoretical framework for evaluating deterministic binary forecasts. This framework evaluates the forecast skill by counting the numbers of times the event was forecasted—or not—and observed—or not—in a contingency table. There are numerous statistics and skill scores that can be computed from the data collected by this method. When applied to a neighborhood, the fraction of grid squares within the neighborhood that contain events (i.e., the modeled and observed values met the threshold criterion) is compared with the total number of grid squares to derive a fractional coverage of events. From this information the FSS is computed, expressing the skill of the model in terms of the spatial scale and threshold value.

In her review of fuzzy verification techniques, Ebert (2008) points out the benefits of this approach in crediting forecasts that are close enough to show skill while providing additional information about the model, which can be used to improve

the model. For example, fuzzy verification provides a measure of the quality of the forecast as a function of spatial scale through the use of neighborhoods. Another example is the ability to relate the skill of the forecast to the value of the threshold and the spatial scale in a way that allows one to identify the scale at which the model shows the desired level of skill. This information allows model developers to determine baseline levels of performance that can be compared with the same information generated following model upgrades to see details about how the upgrade improved the skill. For this study, the FSS was used as the metric of forecast skill.

Jolliffe and Stephenson (2012) discuss the general expectations of model performance arising from the application of fuzzy verification methods to precipitation forecasts. The first expectation is the forecast skill for low-threshold values should exceed that for higher-threshold values. The second is that the skill of the forecast should increase with increasing scale. Ebert (2008), using her case study for precipitation, relates the two by linking the highest skills with low threshold and large spatial scale. These expectations were applied in the analysis phase of this study that involved continuous meteorological variables instead of precipitation, which is a discrete variable, to explore the applicability of fuzzy verification to continuous variable forecasts from the WRE–N using the FSS as the metric.

2. Domain and Model

The ARL WRE–N (Dumais et al. 2004; Dumais et al. 2013) has been designed as a convection-allowing application of the WRF–ARW model (Skamarock et al. 2008) with an observation-nudging FDDA option (Liu et al. 2005; Deng et al. 2009). For this investigation, the WRE–N was configured to run over a multinest set of domains to produce a fine inner mesh with 1.75-km grid spacing and leveraged an external global model for cold-start initial conditions and time-dependent lateral boundary conditions for the outermost nest. Table 1 describes the dimensions for the triple-nested domain. This global model for ARL development and testing has been the National Center for Environmental Prediction's Global Forecast System (GFS) model (EMC 2003). The WRE–N is envisioned to be a rapid-update cycling application of WRF–ARW with FDDA and optimally could refresh itself at intervals up to hourly (dependent on the observation network) (Dumais et al. 2012; Dumais and Reen 2013).

Table 1 WRE–N triple-nested domain dimensions in kilometers

East-west dimension	North-south dimension	Grid spacing
1780	1780	15.75
761	761	5.25
506	506	1.75

For this study, the model runs had a base time of 1200 coordinated universal time (UTC) and produced output for each hour from 1200 to 0600 UTC of the following day for a total of 19 hourly model outputs, which were produced for each of 5 days in February and March 2012. The modeling domains are depicted in Fig. 1.

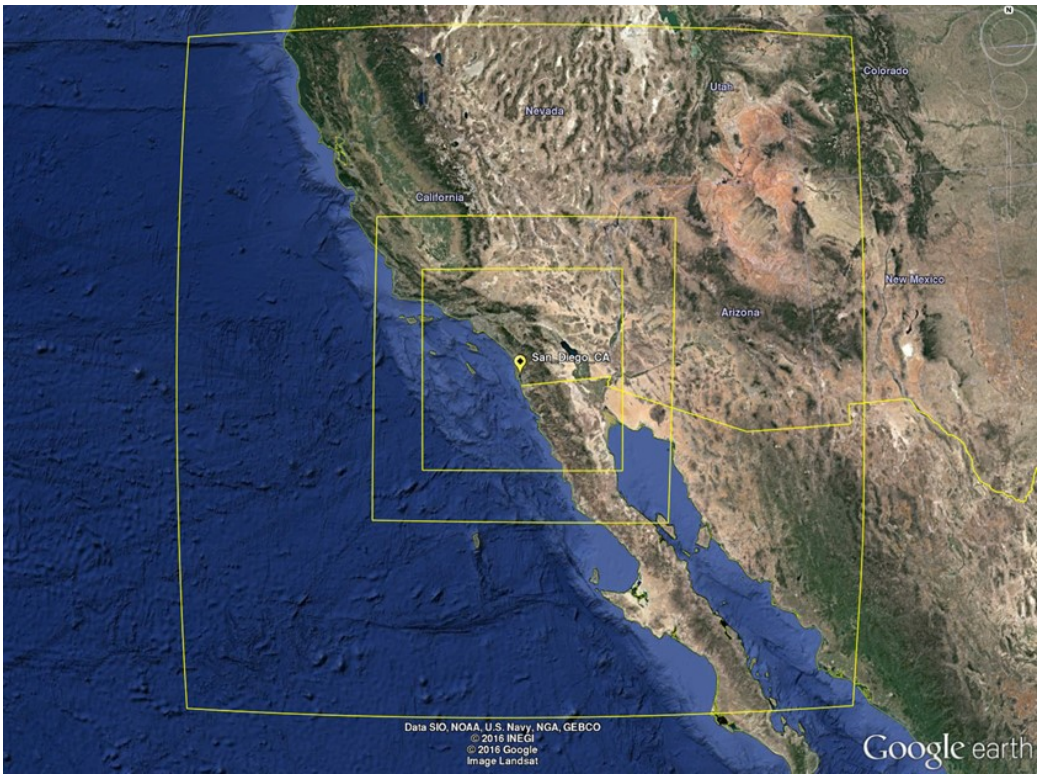


Fig. 1 Triple-nested model domains: domain center points are coincident and are centered near San Diego, CA (Google Earth 2016)

2.1 Observations for Assimilation

The initial conditions were constructed by starting with the GFS data as the first guess for an analysis using observations. Most observations were obtained from the Meteorological Assimilation Data Ingest System (MADIS) (NOAA 2014), except for the Tropospheric Airborne Meteorological Data Reporting (TAMDAR) (Daniels et al. 2006) observations, which were obtained from AirDat, LLC. The MADIS database included standard surface observations, mesonet* surface

*A network of automated meteorological observation stations.

observations, maritime surface observations, wind-profiler measurements, rawinsonde soundings, and Aircraft Communications Addressing and Reporting System (ACARS) data. Use and reject lists were obtained from developers of the RTMA system (De Pondeca et al. 2011), and these were used to filter MADIS mesonet observations. This quality-assurance evaluation is especially important given the greater tendency of mesonet observations to be more poorly sited than other, more standard, surface observations.

The Obsgrid component of WRF was used for quality control of all observations. This included gross-error checks, comparison of observations to a background field (here GFS), and comparison of observations to nearby observations. A modified version of Obsgrid allows for single-level observations such as the TAMDAR and ACARS data to be more effectively compared against the GFS background field. The quality-controlled observations were output in hourly, "little_r"-formatted text files for use as ground-truth data for model assessment. Observation nudging was applied to the model from these same sources for the preforecast period of 1200–1800 UTC (0- through 6-h lead times), followed by 1-h ramping down of the nudging from 1800 to 1900 UTC, during which no new observations are assimilated. The true, free forecast period thus begins at 1800 UTC because no observations after this time are assimilated.

2.2 Parameterizations

For the parameterization of turbulence in WRE–N, a modified version of the Mellor–Yamada–Janjić (MYJ) planetary boundary layer (PBL) (Janjić 1994) scheme was used. This modification decreases the background turbulent kinetic energy and alters the diagnosis of the boundary-layer depth used for model output and data assimilation (Reen et al. 2014). The WRF single-moment, 5-class microphysics parameterization is used on all domains (Hong et al. 2004), while the Kain–Fritsch (Kain 2004) cumulus parameterization is used only on the 15.75-km outer domain. For radiation, the Rapid Radiative Transfer Model (RRTM) parameterization (Mlawer et al. 1997) is used for longwave radiation and the Dudhia (1989) scheme for shortwave radiation. The Noah land-surface model (Chen and Dudhia 2001a, 2001b) is used. Additional references and other details for these parameterization schemes are available from Skamarock et al. (2008). Table 2 lists the WRF configuration settings.

Table 2 WRE–N configuration

Configuration	Y/N?
WRF-ARW V3.4.1	Yes
Obs-nudging FDDA	Yes
Multinest (15.75/5.25/1.75 km)	Yes
MADIS observations (FDDA)	Yes
TAMDAR observations (FDDA)	Yes
Ship/buoy observations (FDDA)	Yes
Filter obs (use/reject) (FDDA)	Yes
RUNWPSPLUS QC (FDDA)	Yes
Obs-nudge rad 120,60,20	Yes
MYJ-PBL scheme (modified)	Yes
WRF, sgl-moment, 5-class mp	Yes
Option 8: microphysics	Yes
End FDDA 360 min	Yes
Kain-Fritsch Cum Param (outer dom)	Yes
RRTM long-wave rad (Mlawer 1997)	Yes
Short-wave rad (Dudhia 1989)	Yes
Noah land surface model	Yes
Fix for nudge to low water vapor	Yes
Model top 10 hectopascal (hPa)	Yes
Feedback on	Yes
Obs weighting function 4E-4	Yes
57 vertical levels	Yes
48-s time step	Yes

2.3 Case-Study Days

The case-study days were selected on the basis of the prevailing synoptic weather conditions over the nested domains. Table 3 provides a short description of these conditions. For this study, only results from analysis of Case 1 data are presented.

Table 3 Synoptic conditions for the case study days considered

Case	Dates (all 2012)	Description
1	February 07–08	Upper-level trough moved onshore, which led to widespread precipitation in the region.
2	February 09–10	Quiescent weather was in place with a 500-hPa ridge centered over central California at 1200 UTC.
3	February 16–17	An upper-level low located near the California–Arizona border with Mexico at 1200 UTC brought precipitation to that portion of the domain. This pattern moved south and east over the course of the day.
4	March 01–02	A weak shortwave trough resulted in precipitation in northern California at the beginning of the period that spread to Nevada, then moved southward and decreased in coverage.
5	March 05–06	Widespread high-level cloudiness due to weak upper-level low pressure but very limited precipitation.

2.4 Observations for Verification

The LAPS gridded observation data sets produced by NOAA–GSD consisted of 12 hourly Gridded Binary–format, edition 2 (GRIB2), files of 2-m above-ground-level (AGL) temperature (TMP), relative humidity (RH), and dew-point temperature (DPT), plus 10-m AGL U-component and V-component winds for the period of 1200–2300 UTC (forecast lead times 0–11) on each of the 5 cases. The output grid used by the LAPS was 289 × 289 with 1.75-km grid spacing. For this study, observations for 1200 UTC in the preforecast (assimilation) period and 1900 UTC in the forecast period (lead times 0 and 7, respectively) were used.

3. Data Preparation Using MET

The model and observational data were preprocessed into the formats required by MET Grid-Stat. The WRE–N model output data were converted from native Network Common Data Form files to hourly Gridded Binary format, edition 1 (GRIB), files by the WRF Unified Post Processor, which destaggers the data onto an Arakawa-A Grid containing 288 × 288 grid points. The hourly GRIB2 observation files on a 289 × 289 grid had to be remapped to the 288 × 288 grid to match that of the WRE–N grid. The NCAR "COPYGB" utility program was used to remap the observations and convert the files to GRIB format (DTC 2016). MET Grid-Stat was used to generate the grid-to-grid, fuzzy verification scores and contingency-table statistics aggregated over the entire 1.75-km WRE–N domain for surface meteorological variables TMP and DPT in degrees Kelvin (K), RH (%), and wind speed (WIND) in meters per second. Grid-Stat applied the specified neighborhood sizes (spatial scales) and thresholds and computed the neighborhood fractional coverage, contingency-table statistics, and skill scores for each spatial scale for 1200 and 1900 UTC for Case 1. The event-coverage threshold used to decide whether the amount of fractional coverage for a neighborhood was considered a "hit" was greater than or equal to 0.5. The variable thresholds were specified using the FORTRAN convention of "GE" to indicate "greater than or equal to" the given threshold value and are shown in Table 4.

Table 4 Thresholds used in MET Grid-Stat

TMP (K)	DPT (K)	RH (%)	WIND (m/s)
265	262	25	2
270	267	40	5
275	272	55	8
280	277	70	11
285	282	85	14
290	. . .	. . .	. . .
295	. . .	. . .	. . .

The neighborhood sizes, in terms of number of grid squares (horizontal direction) and corresponding spatial scales for the 1.75-km grid spacing used in this study, are shown in Table 5.

Table 5 Neighborhood sizes and spatial scales (km) used in MET Grid-Stat

Grid squares	TMP	DPT	RH	WIND
1	1.75	1.75	1.75	1.75
3	5.25	5.25	5.25	5.25
5	8.75	8.75	8.75	8.75
7	12.25	12.25	12.25	12.25
9	15.75	15.75	15.75	15.75
11	19.25	. . .	. . .	. . .
13	22.75	. . .	. . .	. . .
15	26.25	. . .	. . .	. . .
17	29.75	. . .	. . .	. . .

MET Grid-Stat generates many fuzzy-verification skill scores and traditional contingency-table statistics. Of these, Table 6 lists those that were output initially for this study.

Table 6 Initial Grid-Stat fuzzy-verification skill scores and contingency-table statistics

Score/statistic	Description
FSS	Fractions skill score
BASER	Base rate
FMEAN	Mean forecast value
PODY	Probability of detection—hit rate
FAR	False-alarm ratio
FBIAS	Frequency bias
CSI	Critical success index
GSS	Gilbert skill score
ACC	Accuracy
MCTC	Multicategory contingency table counts

The Grid-Stat output text files were ingested into Microsoft Excel spreadsheets, which were used to generate tabular and graphical displays showing how the scores and statistics, aggregated over the entire domain, vary with threshold value at different spatial scales. Text files containing the MCTC data were ingested into spreadsheets and plotted to show the distribution of counts of the forecast variable in various bins or ranges and the corresponding counts occurring in the same bins of the observed variable. This information is used to study model performance over discrete ranges of the variables. For this study, the analysis considered only FSS for the variables of 2-m-AGL TMP and RH and 10-m-AGL WIND, reducing the burden of analysis that comes when considering numerous scores and statistics during a preliminary evaluation of the applicability of the FSS in assessing the baseline performance of WRE–N and in tracking changes in performance following model upgrades.

4. Data Analysis of MET Grid-Stat Fuzzy Verification Results

4.1 Fuzzy Verification Statistics and Scores

Traditional contingency table statistics are defined by a ratio of counts determined using a 2 × 2 contingency table. Table 7 shows the contingency table with notation consistent with the formulae for the scores and statistics as implemented in the MET (NCAR 2013).

Table 7 The 2 × 2 contingency table from the MET User's Guide 4.1

2x2 contingency table in terms of counts. The n_{ij} values in the table represent the counts in each forecast-observation category, where i represents the forecast and j represents the observations. The "." symbols in the total cells represent sums across categories.

Forecast	Observation		Total
	o = 1 (e.g., "Yes")	**o = 0 (e.g., "No")**	
f = 1 (e.g., "Yes")	n_{11}	n_{10}	$n_{1\cdot} = n_{11} + n_{10}$
f = 0 (e.g., "No")	n_{01}	n_{00}	$n_{0\cdot} = n_{01} + n_{00}$
Total	$n_{\cdot 1} = n_{11} + n_{01}$	$n_{\cdot 0} = n_{10} + n_{00}$	$T = n_{11} + n_{10} + n_{01} + n_{00}$

The counts, n_{11}, n_{10}, n_{01}, and n_{00}, are sometimes called the "Hits", "False alarms", "Misses", and "Correct rejections", respectively.

By dividing the counts in the cells by the overall total, T, the joint proportions, p_{11}, p_{10}, p_{01}, and p_{00} can be computed. Note that $p_{11} + p_{10} + p_{01} + p_{00} = 1$. Similarly, if the counts are divided by the row (column) totals, conditional proportions, based on the forecasts (observations) can be computed.

The FSS score (Eq. 1) is computed as described by Mittermaier et al. (2013):

$$\text{FSS} = 1 - \frac{FBS}{FBS_{worst}}, \tag{1}$$

where Fractions Brier Score (FBS) (Eq. 2) is defined as:

$$FBS = \frac{1}{N}\sum_{i=1}^{N}(O_i - F_i)^2, \tag{2}$$

where N is the number of neighborhoods, O_i is the proportion of grid boxes within an observed neighborhood containing events, and F_i is the proportion of grid boxes in a forecast neighborhood containing events.

FBS_{worst} is the worst possible FBS, which is the case when there is no coincidence between the forecast and observed events. An FSS value of 0 indicates a forecast with no skill and a value of 1 indicates a forecast with perfect skill.

4.2 Apply FSS as a Metric for WRE–N Performance

The analysis of the FSS values focuses on the forecast accuracy as defined by FSS as well as the degree to which the trend of skill, as a function of threshold value and spatial scale, follows the 2 expectations described in Section 1:

- Skill increases with decreasing threshold value
- Skill increases with increasing spatial scale

The ranges of the variables over the 1.75-km WRE–N domain at the 7-h lead time 1900 UTC were used to establish the bounds within which the threshold values were selected. Initially, the range was divided into 4 intervals of equal size to determine the initial threshold values, with the highest value selected such that it would delineate an interval containing extreme values near the maximum value of the range. This simulated a situation similar to that which occurs when the MyWIDA TDA, which uses absolute thresholds, indicates an unfavorable impact based on forecast values exceeding the worst system or mission threshold.

Initial analyses of the FSS for TMP revealed a pattern of variability of the values with threshold that varied little with spatial scale. It was decided that the range of spatial scales should be expanded to include larger neighborhood sizes, capturing a stronger signal of the dependence of forecast skill on spatial scale. During a review of some early results, one reviewer suggested that expanding the number of thresholds and scales might reveal additional information about the relationship between the threshold and the scale (Smith 2016). Analysis of the results produced with this expanded range of scales and larger number of thresholds failed to show a clear relationship between skill and scale. It was decided that for the other variables a reduced or nominal range of scales and number of thresholds would be sufficient to represent the full range of score variation with spatial scale for this study, which is focused on illustrating the applicability of the FSS as a metric for tracking model performance rather than presenting verification results. Scores were computed for a WRE–N valid time that fell in hour 7 of the forecast period to provide examples to characterize the forecast skill in terms of threshold value and spatial scale.

Tables and plots of FSS versus threshold at various spatial scales for the variables TMP, RH, and WIND are now presented with an analysis of their characteristic features and their possible implications regarding model skill. In addition, 2-D maps of model TMP, RH, and WIND fields from the WRE–N forecast valid at 1900 UTC were generated using a solid-color assignment to illustrate the spatial distribution and domain coverage of the variable where the value equals or exceeds

each threshold. In further discussions regarding features in these maps, the term "object" will be used to refer to the areas shown where the value of the variable meets or exceeds the threshold. (Statements made regarding model performance only illustrate how an indication of performance can be gained from the FSS taken alone.)

Generally speaking, the FSS scores in Table 8 confirm what we expect; that is, the scores increase (decrease) with decreasing (increasing) threshold and increasing (decreasing) neighborhood size. However, the change of scores as a function of threshold is rather small below 285 K, while the score dropped by approximately 50% when the threshold increased to 290 K and even ceased to exist for a threshold of 295 K. This suggests that something fundamentally changed when the threshold changed from 285 to 290 K. We attempt to explain what caused this dramatic drop of FSS scores starting at 285 K later in this section.

Table 8 FSS for WRE–N TMP valid at 1900 UTC for Case 1

2-m-AGL TMP (K) FSS 19Z (07 Feb 2012)							
Spatial scale (km)	**Threshold (K)**						
	265	270	275	280	285	290	295
29.75	1	1	0.99994	0.99947	0.99315	0.54137	NaN
26.25	1	1	0.99993	0.99938	0.99248	0.53778	NaN
22.75	1	1	0.99991	0.99928	0.9917	0.53349	NaN
19.25	1	1	0.99989	0.99915	0.99075	0.5283	NaN
15.75	1	1	0.99987	0.99898	0.98954	0.52204	NaN
12.25	1	0.99999	0.99984	0.99874	0.98792	0.51446	NaN
8.75	1	0.99999	0.99979	0.99834	0.98568	0.50485	NaN
5.25	1	0.99997	0.99968	0.9976	0.9821	0.49159	NaN
1.75	1	0.99989	0.99916	0.9949	0.97213	0.46108	NaN

The other notable feature from Table 8 is the fact that the FSS scores do not change much as a function of neighborhood size even though the general trend of increasing scores with increasing neighborhood size still holds. There are 2 possible reasons that work together to cause this; one is related to neighborhood size versus object size, and the other is related to neighborhood size versus displacement errors of forecast objects. Again, we will defer details of this discussion to the following paragraphs when we show the forecast object size for various thresholds.

Figure 2 is a display of FSS versus threshold value for the expanded range of spatial scales for TMP at valid time 1900 UTC (lead time = 7 h) for Case 1.

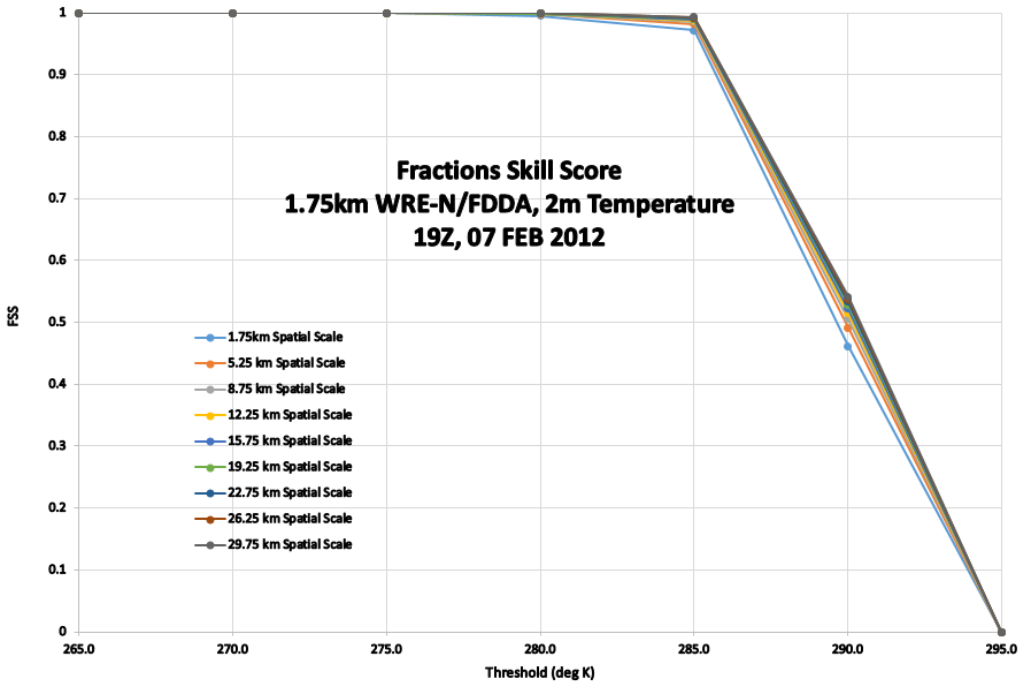
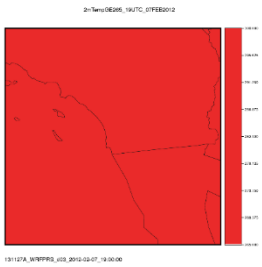
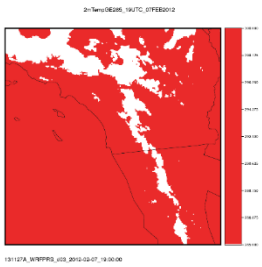


Fig. 2 FSS vs. threshold for a range of spatial scales for 2-m-AGL TMP at 1900 UTC for Case 1

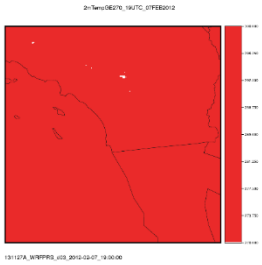
At 1900 UTC, the trend of decreasing FSS with increasing threshold value shows good skill for the WRE–N for threshold values of 265–285 K, but the skill decreases markedly for higher threshold values, which is the expected trend overall. Two features from Fig. 2 are worth noting. First, there is little spread of FSS scores for different neighborhood sizes, as demonstrated by the curves almost coinciding for all threshold values, except for 290 K, where there is a small spread in FSS values, with increasing FSS value for increasing spatial scales as we would expect. Second, the FSS scores are fairly high and do not change much for thresholds up to 285 K; then they start to drop dramatically and basically become zero for a threshold of 295 K. The answer for the behavior for FSS scores in Fig. 2 will be explored using data from Fig. 3, which maps the spatial distribution of WRE–N TMP, color-shaded to depict the spatial distribution (object) of the variable defined where its value equals or exceeds each threshold at forecast valid time 1900 UTC for Case 1.



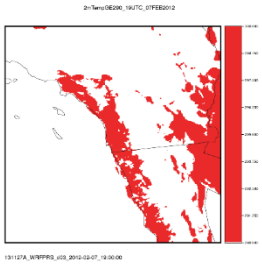
TMP GE 265 K



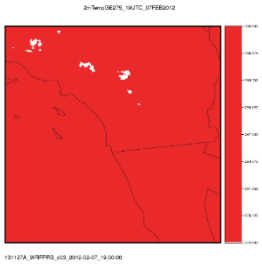
TMP GE 285 K



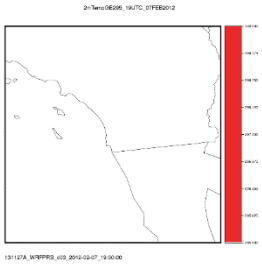
TMP GE 270 K



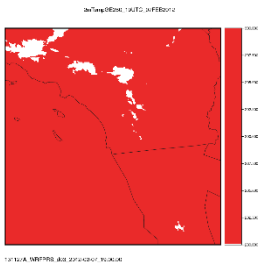
TMP GE 290 K



TMP GE 275 K



TMP GE 295 K



TMP GE 280 K

Fig. 3 Map of WRE–N 2-m-AGL TMP, GE the 7 threshold values at 1900 UTC for Case 1

The spatial extent of the TMP object as defined by the first threshold (265 K) covers the entire domain. For thresholds of GE 270 K, the objects occupy a decreasing amount of the domain area. At GE 295 K, no object is present because no WRE–N TMP value equals or exceeds the threshold. The threshold value for which forecast skill, as indicated by FSS, begins a sharp decline (285 K) does coincide with the dramatic drop in object size starting at 285 K. When the object size becomes smaller, it is more difficult to obtain a perfect match between forecast and observed objects (Cai and Dumais, 2015), which causes a decrease in FSS scores. On the other hand, because the neighborhood size (1.75–29.75 km) is relatively small compared with the object size for all the thresholds, and on top of that it may also be much smaller than the displacement error of forecast object based on Fig. 3. Thus, it is not surprising that all the curves for different neighborhood size in Fig.2 are closely tracking each other. A good test for this explanation would be to use a much larger range of neighborhood size to see if the expected spread in FSS will appear.

The possible relationship between object size and FSS scores for TMP may have implications for assessing the ability of the WRE–N to predict objects that in turn impacts the input data used by MyWIDA. It is very important to understand that forecast skills are very much affected by the thresholds used; therefore, it is imperative to investigate what size of objects we normally obtain by applying the actual thresholds employed in MyWIDA so that reasonable expectations of model performance can be obtained. Certainly, analysis of more data is needed to confirm this apparent loss of skill when forecast objects are defined by the upper part of the range of the TMP thresholds.

Similar to Table 8, the RH FSS scores in Table 9 also confirmed the general trends of FSS scores as a function of threshold and neighborhood size. Again, the scores do not change much as a function of neighborhood size, but they do show a dramatic drop of FSS scores from 70% to 85% RH threshold. We believe a similar argument, as we proposed earlier discussing Table 8 results, which involves object size and relative small ratio of neighborhood size versus object size and/or neighborhood size versus forecast displacement error, could be invoked to explain the FSS scores shown in Table 8 in the following paragraphs.

Table 9 FSS for WRE–N RH valid at 1900 UTC for Case 1

Spatial scale (km)	2-m-AGL RH (%) FSS 19Z (07 Feb 2012)				
	Threshold (%)				
	25	40	55	70	85
15.75	0.9373	0.93085	0.96928	0.94472	0.68245
12.25	0.9336	0.92641	0.9648	0.9401	0.68248
8.75	0.92913	0.92147	0.95956	0.93508	0.68201
5.25	0.9234	0.91553	0.95316	0.92915	0.68063
1.75	0.91309	0.90446	0.94152	0.91875	0.67523

Figure 4 displays FSS versus threshold value for the nominal range of spatial scales for WRE–N RH at the forecast valid time 1900 UTC (lead time = 7 h) for Case 1.

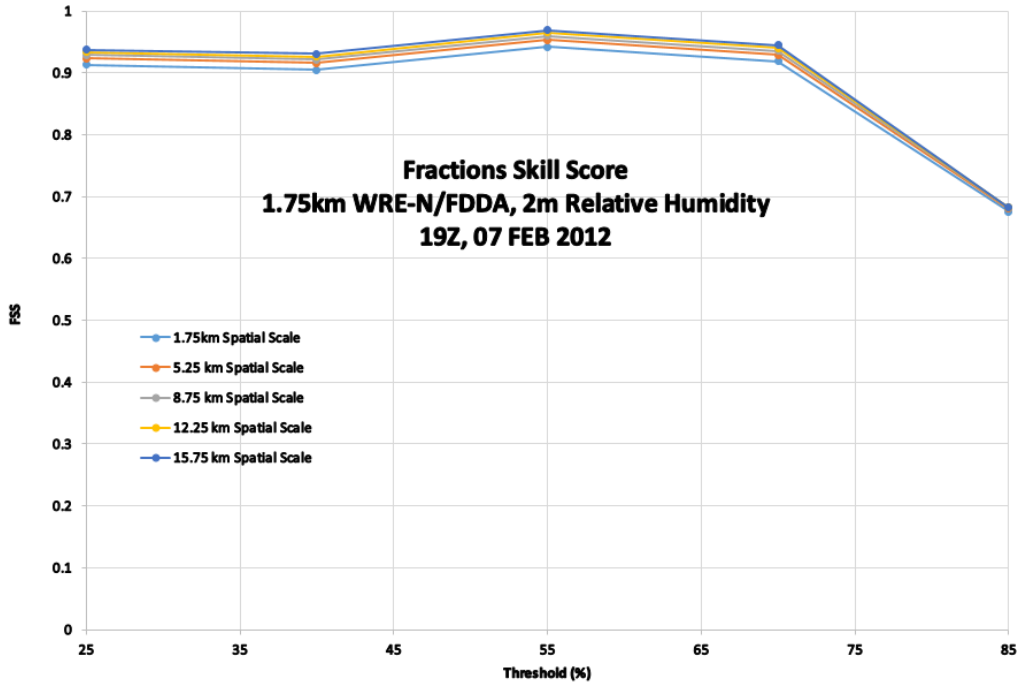


Fig. 4 FSS vs. threshold for a range of spatial scales for WRE–N 2-m-AGL RH at 1900 UTC for Case 1

At 1900 UTC, the WRE–N shows good scores over a wider range of thresholds. The FSS scores are good between 25% and 70%, then decrease with threshold after 70%, though never as low as that for TMP after its performance drops at higher threshold values. Overall, the plot shows the expected trend of FSS decreasing with increasing threshold value. There is no significant difference in FSS scores at a fixed threshold value with spatial scale, but the expected trend still holds.

Figure 5 shows the spatial distribution of WRE–N RH color-shaded to depict the spatial distribution of the variable (object) defined where its value equals or exceeds each threshold at forecast valid time 1900 UTC for Case 1.

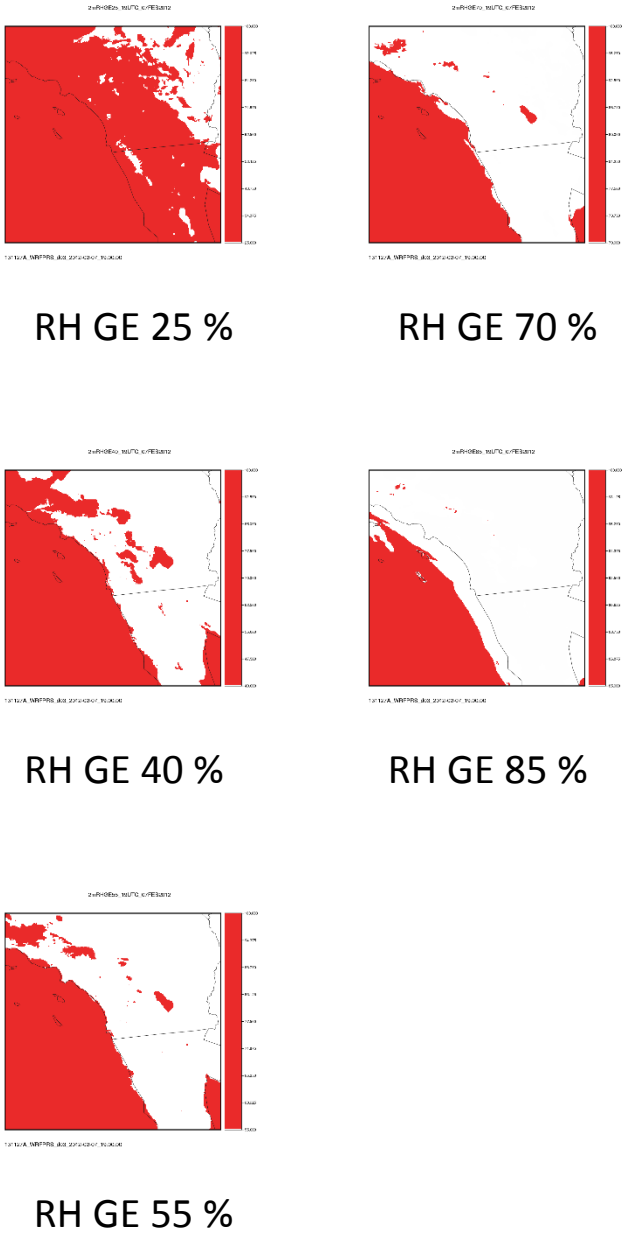


Fig. 5 Map of WRE–N 2-m-AGL RH, GE the 5 threshold values at 1900 UTC for Case 1

The spatial extent of the RH objects for all 5 thresholds covers a large portion of the domain between 25% and 85%. Object size progressively decreases to less than half of the domain at 85%. The threshold value for which forecast skill, as indicated by FSS, begins a decline (70%) that coincides with the same threshold value for which the object is roughly half of the size of the domain. However, the RH object

size does not decrease that much compared with the TMP object size at 285 K. This explains why the corresponding RH FSS score also does not decrease as much as that for TMP (Fig. 2). The possible relationship between object size and FSS for RH may have implications for assessing the ability of the WRE–N to predict objects that in turn impacts the input data used by MyWIDA. One factor that may contribute to this decrease in skill is the smaller objects, which are defined at higher-threshold values. Matching for small objects between the forecast and the observations tends to be difficult because it requires a smaller displacement error. Analysis of more data is needed to confirm this apparent loss of skill when forecasting objects defined by the upper part of the range of the RH thresholds.

Table 10 presents the FSS scores for the WRE–N WIND forecast valid at 1900 UTC.

Table 10 FSS for WRE–N WIND valid at 1900 UTC for Case 1

Spatial scale (km)	10-m-AGL WIND (m/s) FSS 19Z 07 FEB 2012				
	Threshold (m/s)				
	2	5	8	11	14
15.75	0.90273	0.85515	0.83594	0.40038	0.09435
12.25	0.8979	0.84752	0.82419	0.37624	0.09087
8.75	0.89212	0.83828	0.81026	0.35205	0.08677
5.25	0.88457	0.8265	0.79257	0.32677	0.08594
1.75	0.87041	0.80591	0.76244	0.29207	0.08022

Similar to Tables 8 and 9, Table 10 also confirms the general trends of FSS score as a function of threshold and neighborhood size. The FSS scores for different neighborhood size are closely compacted, just like TMP and RH FSS scores shown earlier, which indicates that the impact of neighborhood size on FSS scores used in this study is small. The dramatic drop in FSS scores starting at 8 m/s suggests something fundamentally changed, which we will address later.

Figure 6 displays FSS versus threshold value for the nominal range of spatial scales for WRE–N WIND at the forecast valid time 1900 UTC (lead time = 7 h) for Case 1.

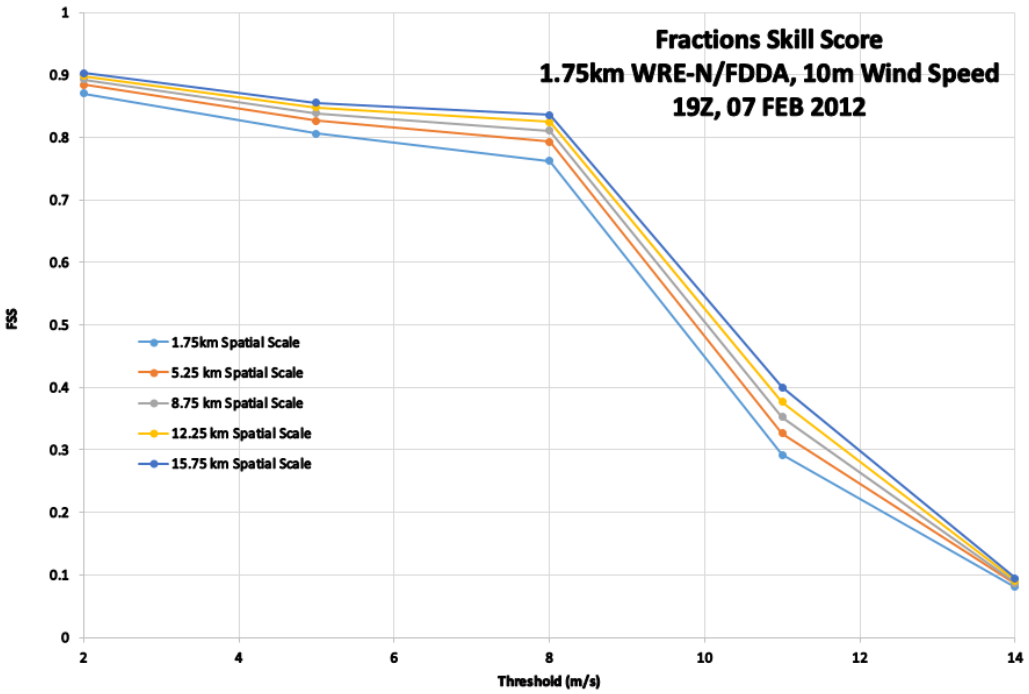


Fig. 6 FSS vs. threshold for a range of spatial scales for WRE–N 10-m-AGL WIND at 1900 UTC for Case 1

The performance of the WRE–N over the entire range of threshold values is fair. The overall trend of decreasing FSS with increasing threshold value is as expected. However, the overall skill for WIND is less than for TMP and RH, which is consistent with other model evaluation results. The FSS does not drop significantly until after the threshold reaches a value of 8 m/s. There is no significant difference in FSS at a fixed threshold value among the various spatial scales.

Figure 7 depicts the spatial distribution of model WIND color-shaded to show the spatial distribution of the variable (object) defined where its value equals or exceeds each threshold at forecast valid time 1900 UTC for Case 1.

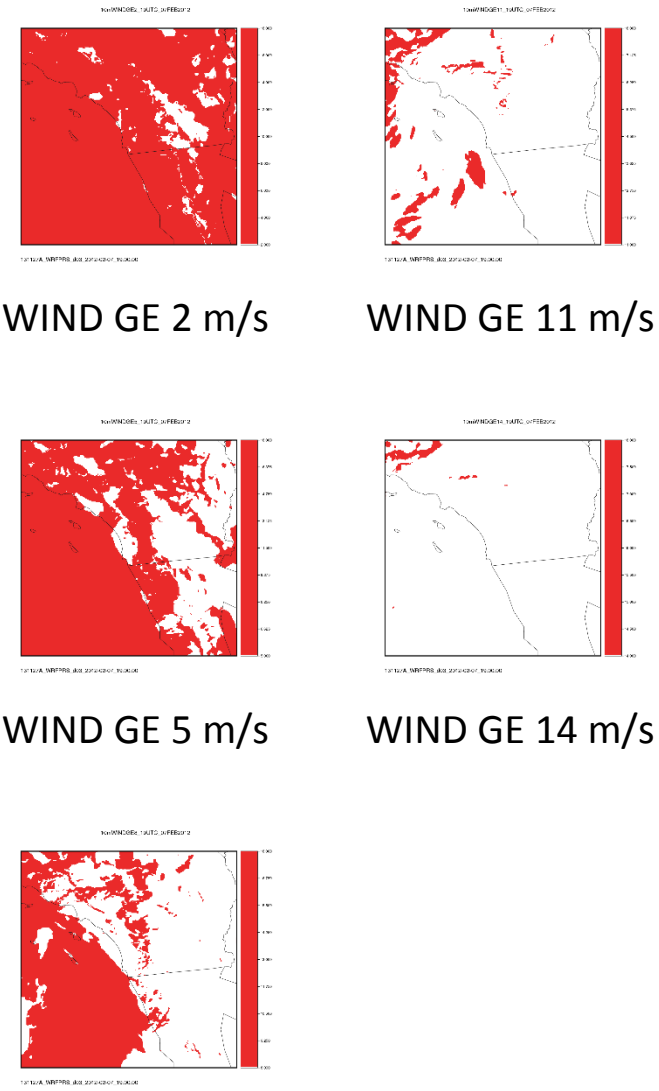


Fig. 7 Map of WRE–N 10-m-AGL WIND, GE the 5 threshold values at 1900 UTC for Case 1

The spatial extent of the WIND object, as defined by the threshold for all 5 thresholds, covers a large portion of the domain at 2 m/s and decreases in size to a very small portion at the highest threshold value of 14 m/s. The threshold value after which forecast skill, as indicated by FSS, begins a sharp decline (8 m/s) coincides with the same threshold value for which the object size decreased dramatically. The possible relationship between object size and FSS for WIND may have implications for assessing the ability of the WRE–N to predict objects that in turn impact the input data used by MyWIDA. Similar to TMP and RH, one factor that may contribute to this decrease in skill is the smaller objects, which are defined

at higher threshold values. Matching for small objects between the forecast and the observations tends to be difficult because it requires a smaller displacement error. Analysis of more data is needed to confirm this apparent loss of skill when forecasting WIND objects defined over the higher portion of the range of WIND thresholds.

Fuzzy verification methods such as FSS provide a simple, straightforward way to compare different forecasts as a function of threshold and neighborhood size. It gives credit for slightly misplaced forecasts by comparing fractional coverage within a predetermined neighborhood around the grid point where the verification is performed. Therefore, FSS is suitable for evaluating high-resolution NWP model forecasts such as WRE–N, and it is widely used by many researchers in the modeling community. For this reason, the FSS is particularly well-suited for evaluating the effectiveness of changes made to the model during the development process. The dependency of FSS on threshold value and neighborhood size provides significant information that can be compared before and after model upgrades that can help determine how and if the change achieved the desired effect.

This report documented our first attempt to evaluate continuous variables as a categorical forecast by applying a threshold using the FSS. As demonstrated in the examples presented in this section, we found that 1) the neighborhood size does not affect the FSS scores too much and 2) there is a certain threshold for each variable, and when that threshold is exceeded, a dramatic drop of FSS scores occurred. One possible reason for the former is the neighborhood size, even as large as approximately 30 km, is still relatively smaller than the typical object size determined by the threshold used. Besides, the average displacement error of forecast and observed objects may be much larger than the maximum neighborhood size of about 30 km, therefore, increasing the neighborhood size up to about 30 km would not increase the FSS scores significantly.

As for the reason why the FSS scores always drop dramatically after certain thresholds for all cases, we believe it is probably related to object size. A careful examination of object size reveals that the large drop in FSS scores is always coincident with the large drop in forecast object size, which makes matching of forecast and observed objects more difficult. Cai (2016) reviewed the results presented by Raby (2016), who discussed similar results and attributed the lack of forecast skill at high thresholds to small object size. For the same displacement error, large objects will result in a significant number of hits and decreases in the number of misses compared with smaller objects, which result in a significant decrease in the number of hits and increases in the number of misses. The impact of this situation is to lower the FSS for small objects and raise it for large objects.

These findings have profound implications on TDAs such as MyWIDA, which is threshold-based. It is thus imperative to find out the typical object size created by MyWIDA thresholds in a particular region using climatological conditions so that reasonable expectations of model performance can be anticipated.

5. Summary and Conclusion

The FSS method of fuzzy verification, supported by output from the MET Grid-Stat tool, can be applied to the assessment of high-resolution WRE–N model forecasts (Ebert 2008). Furthermore, the tool employs categorical verification techniques that involve the application of threshold values to the spatial fields of continuous meteorological variables, offering the added benefit of a unique type of verification of the WRE–N's ability to simulate objects defined by these thresholds—analogous to objects depicting areas of marginal and unfavorable weather impacts on Army missions and systems, which are the product of the MyWIDA TDA.

It was found that a nontraditional fuzzy-verification technique that uses the FSS offers a more robust approach to assess the ability of the model to predict objects defined by the application of a threshold to a spatial forecast of a continuous variable. This study demonstrated the applicability of the MET Grid-Stat tool for computing the FSS for a range of threshold values and neighborhood sizes that establish the baseline performance of the WRE–N. As the model undergoes upgrades, the FSS can be recalculated to reveal any differences in skill as a function of threshold value and spatial scale. This way, the model developers can determine if there is a trend of increasing or decreasing skill at specified scales and thresholds that may indicate that the implemented change is having a beneficial or adverse effect. Mittermaier et al. (2013) noted that the FSS can be used to determine if decreasing the grid spacing results in improved skill.

This study's results suggest that, for the range of spatial scales and the number of thresholds used, the expected trend of decreasing model skill with increasing threshold value was confirmed; however, there is no significant difference using the range of spatial scales as those tested in this report, although the trend of increasing skills with increasing neighborhood size still holds. A possible explanation of these observations may be that the neighborhood sizes used in this study are rather small compared with the object size and/or the displacement error of the forecast objects. Expanding the neighborhood size would be a good test to confirm this hypothesis.

Another discovery of this study was finding a threshold that marks the steep drop of FSS scores for continuous variables such as surface temperature, relative humidity, and wind speed. It was found that this large drop of forecast skills is usually coincident with the large drop of object size determined by a particular threshold. Since MyWIDA also uses thresholds applied to the forecast input data to identify potential hazardous regions for Army operations, it is imperative to understand how the size of the object determined by MyWIDA thresholds affects model performance. Future work should include further investigation in this area so that a reasonable expectation of WRE–N performance for certain MyWIDA thresholds can be inferred.

To further improve assessments of the predictability of objects, Raby and Cai (2016) recommended a more rigorous approach that requires the generation of larger data sets of forecast output and gridded observations so that statistically significant results can be obtained. This will be important when verifying the modeled objects defined at higher thresholds, particularly when WRE–N model output is used to predict the more critical unfavorable-weather impacts on Army systems and missions using MyWIDA.

6. References

Brandt J, Dawson L, Johnson J, Kirby S, Marlin D, Sauter D, Shirkey R, Swanson J, Szymber R, Zeng S. Second-generation weather impacts decision aid applications and web services overview. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 July. Report No.: ARL-TR-6525.

Cai H. US Army Research Laboratory, White Sands Missile Range, NM. Personal communication; 2016 Mar 15.

Cai H, Dumais RE. Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. Weather Forecast. 2015;30:1451–1468.

Casati B, Wilson LJ, Stephenson DB, Nurmi P, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown BG, Mason S. Forecast verification: current status and future directions. Meteorol Appl. 2008;15(1):3–18.

Chen F, Dudhia J. Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part II: preliminary model validation. Mon Weather Rev. 2001a;129:587–604.

Chen F, Dudhia J. Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon Weather Rev. 2001b;129:569–585.

Daniels TS, Moninger WR, Mamrosh RD. Tropospheric airborne meteorological data reporting (TAMDAR) overview. Preprints for the 10th Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface; 2016 Sep 1. Atlanta (GA): American Meteorological Society [accessed 2016 Aug 2]. http://ams.confex.com/ams/pdfpapers/104773.pdf.

De Pondeca Manuel SFV, Manikin G, DiMego G, Benjamin S, Parrish D, Purser RJ, Wu WS, Horel J, Myrick D, Lin Y, et al. The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: current status and development. Weather Forecast. 2011;26:593–612.

Deng A, Stauffer D, Gaudet B, Dudhia J, Hacker J, Bruyere C, Wu W, Vandenberghe F, Liu Y, Bourgeois A. Update on the WRF-ARW end-to-end multi-scale FDDA system. Presented at the 10th WRF Users' Workshop, National Center for Atmospheric Research; 2009 Jun 23–26; Boulder, CO.

[DTC] Developmental Testbed Center. MET online tutorial for METv3.0: COPYGB functionality. Boulder (CO): National Oceanic and Atmospheric Administration [accessed 2016 Jul 27]. http://www.dtcenter.org/met/users/support/online_tutorial/METv3.0/copygb/index.php.

Dudhia J. Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. J Atmos Sci. 1989;46:3077–3107.

Dumais RE, Reen BP. Data assimilation techniques for rapidly relocatable weather research and forecasting modeling. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 June. Report No.: ARL-TN-0546.

Dumais R, Kirby S, Flanigan R. Implementation of the WRF four-dimensional data assimilation method of observation nudging for use as an ARL weather running estimate-nowcast. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 June. Report No.: ARL-TR-6485.

Dumais RE Jr., Henmi T, Passner J, Jameson T, Haines P, Knapp D. A mesoscale modeling system developed for the US Army. White Sands Missile Range (NM): Army Research Laboratory (US); 2004. Report No.: ARL-TR-3183.

Dumais RE, Raby JW, Wang Y, Raby YR, Knapp, D. Performance assessment of the three-dimensional wind field weather running estimate–nowcast and the three-dimensional wind field Air Force Weather Agency weather research and forecasting wind forecasts. White Sands Missile Range (NM): Army Research Laboratory (US); 2012 Dec. Report No.: ARL-TN-0514.

Ebert E. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. Meteorol Appl. 2008;15:51–64.

Ebert E et al. Progress and challenges in forecast verification. Meteorol Appl. 2013;20(2):130–139.

[EMC] Environmental Modeling Center. The GFS atmospheric model. Washington (DC): National Weather Service/National Centers for Environmental Prediction; 2003 Nov. NCEP Office Note No.: 442.

Google Earth. Mountain View (CA); 2016 [accessed 2016 Aug 24]. http://maps.google.com/help/terms_maps.html.

Hong SY, Dudhia J, Chen SH. A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. Mon Weather Rev. 2004;132:103–120.

Janjić ZI. The step-mountain eta coordinate model: further developments of the convection, viscous sublayer, and turbulence closure schemes. Mon Weather Rev. 1994;122:927–945.

Jolliffe IT, Stephenson DB. Forecast verification: a practitioner's guide in atmospheric science. 2nd ed. Hoboken (NJ): John Wiley and Sons; 2012.

Kain JS. The Kain-Fritsch convective parameterization: an update. J App Meteorol. 2004;43:170–181.

Liu Y, Bourgeois A, Warner T, Swerdlin S, Hacker J. Implementation of observation-nudging based FDDA into WRF for supporting ATEC test operations. Presented at the 6th WRF/15th MM5 Users' Workshop, National Center for Atmospheric Research; 2005 Jun 27–30; Boulder, CO.

Mittermaier M, Roberts N, Thompson SA. A long-term assessment of precipitation forecast skill using the fractions skill score. Meteorol Appl. 2013;20:176–186. doi:10.1002/met.296.

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. J Geophys Res-Atmos. 1997;102:16663–16682.

[NCAR] National Center for Atmospheric Research. Model evaluation tools version 4.1 (METv4.1), user's guide 4.1. Boulder (CO): NCAR; 2013 May.

[NOAA] Meteorological assimilation data ingest system (MADIS). College Park (MD): National Oceanic and Atmospheric Administration [accessed 2016 July 27]. http://madis.noaa.gov.

[NRC] National Research Council. When weather matters: science and service to meet critical societal needs. Washington (DC): The National Academies Press; 2010.

Raby JW. Application of a fuzzy verification technique for assessment of the weather running estimate–nowcast (WRE–N) model. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Oct. Report No.: ARL-TR-7849.

Raby JW, Cai H. Verification of spatial forecasts of continuous meteorological variables using categorical and object-based methods. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Aug. Report No.: ARL-TR-7751.

Reen BP, Schmehl KJ, Young GS, Lee JA, Haupt SE, Stauffer DR. Uncertainty in contaminant concentration fields resulting from atmospheric boundary layer depth uncertainty. J Appl Meteorol Clim. 2014;53:2610–2626.

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang X-Y, Wang W, Powers JG. A description of the advanced research WRF version 3. Boulder (CO): National Center for Atmospheric Research; 2008 Jun. NCAR Technical Note No.: TN–475+STR.

Smith J. Personal communication. White Sands Missile Range, NM, US Army Research Laboratory, 2016 Mar 14.

Stauffer DR, Seaman NL. Multiscale four-dimensional data assimilation. J App Meteorol. 1994;33:416–434.

Vaucher G, Raby J. Assessing high-resolution weather research and forecasting (WRF) forecasts using an object-based diagnostic evaluation. White Sands Missile Range (NM): Army Research Laboratory (US); 2014 Feb. Report No.: ARL-TR-6843.

Wilks DS. Statistical methods in the atmospheric sciences. 3rd ed. Oxford (England): Academic Press; 2011.

List of Symbols, Abbreviations, and Acronyms

2-D	2-dimensional
ACARS	Aircraft Communications, Addressing, and Reporting System
AGL	above ground level
ARL	US Army Research Laboratory
ARW	Advanced Research Weather Research and Forecasting
DPT	dew-point temperature
FBS	Fractions Brier Score
FDDA	Four-Dimensional Data Assimilation
FSS	Fractions Skill Score
GE	greater than or equal to
GFS	Global Forecast System
GRIB	Gridded Binary format Edition 1
GRIB2	Gridded Binary format Edition 2
GSD	Global Systems Division
hPa	hectopascal
K	degrees Kelvin
LAPS	Local Analysis and Prediction System
MADIS	Meteorological Assimilation Data Ingest System
MCTC	Multicategory Contingency Table Count
MET	Model Evaluation Tool
MODE	Method for Object-based Diagnostic Evaluation
MYJ	Mellor–Yamada–Janjic
MyWIDA	My Weather Impacts Decision Aid
NCAR	National Center for Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration

NWP	Numerical Weather Prediction
PBL	planetary boundary layer
RH	relative humidity
RRTM	Rapid Radiative Transfer Model
RTMA	Real-Time Mesoscale Analysis
TAMDAR	Tropospheric Airborne Meteorological Data Reporting
TDA	tactical decision aid
TMP	temperature
UTC	coordinated universal time
WIND	wind speed
WRE–N	Weather Running Estimate–Nowcast
WRF	Weather Research and Forecasting
WRF–ARW	Weather Research and Forecasting, Advanced Research WRF

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIR ARL
(PDF) IMAL HRA
RECORDS MGMT
RDRL DCL
TECH LIB

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 US NAVY RSRCH LAB
(PDF) DR J MCLAY

1 US AIR FORCE 557TH WEATHER WING
(PDF) R CRAIG

1 DCGS-A WEATHER EET LEAD
(PDF) J CARROLL

3 UCAR
(PDF) T FOWLER
J H GOTWAY
B BROWN

1 USAICOE
(PDF) J STALEY

13 ARL
(PDF) RDRL CIE
P CLARK
T JAMESON
RDRL CIE D
D KNAPP
S O'BRIEN
J JOHNSON
G VAUCHER
RDRL CIE M
H CAI
J SMITH
J PASSNER
R PENC
R DUMAIS
B REEN
B MACCALL

INTENTIONALLY LEFT BLANK.